ROBUST SPEECH RECOGNITION USING MISSING DATA TECHNIQUES IN THE PROSPECT DOMAIN AND FUZZY MASKS

Maarten Van Segbroeck[*]*, Hugo Van hamme*

Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{maarten.vansegbroeck,hugo.vanhamme}@esat.kuleuven.be

ABSTRACT

Missing data theory (MDT) has been applied to handle the problem of noise-robust speech recognition. Conventional MDT-systems require acoustic models that are expressed in the log-spectral rather than in the cepstral domain, which leads to a loss in accuracy. Therefore, we have already introduced a MDT-technique that can be applied in any feature domain that is a linear transform of log-spectra. This MDT-system requires hard decisions about the reliability of each spectral component. When computed from noisy data, misclassification errors in the mask are hardly unavoidable and the recognition rate will significantly degrade. The risk of misclassifications can be reduced by estimating a probability that the component is reliable, e.g. a fuzzy mask. In this paper, we extend our MDT-system to be applied in the probabilistic decision framework. Experiments on the Aurora2 database demonstrate a further increase in recognition accuracy, especially at low SNRs.

Index Terms— speech recognition, noise robustness, missing data techniques, fuzzy masks

1. INTRODUCTION

Additive noise leads to a decrease in performance of speech recognition systems due to the mismatch between the speech models (obtained in clean conditions) and the statistics of speech in the noisy test conditions. A MDT-based recognizer will handle this noise robustness problem by adding two import modifications to a speech recognizer. In the front-end, the missing data detector (MDD) will decide for each time-frequency cell whether it is dominated by speech or by noise. If a hard decision is made, the missing data mask will indicate that the cell is either completely reliable or else completely missing. The accuracy of such a binary mask is very crucial since mask estimation errors will cause a significant degradation in recognition performance. Previous studies ([1], [2] and [3]) have shown that the MDT-based recognizer achieves better results when fuzzy (or soft) masks are used which represent the probability that a spectral component is reliable.

A second modification in the MDT-recognizer needs to be made during the evaluation of the acoustic model, e.g. the fact that some of the data are missing should be taken into account. For reasons of accuracy, the ASR-system operates in a domain that is a linear transformation of log-spectra, such as cepstra. However, conventional MDT-techniques like bounded marginalization [4], [5], or the imputation techniques of [6], [7], rely on GMMs with diagonal covariance expressed in the log-spectral domain. Therefore, an alternative MDT-technique was introduced in [8] where the spectral representation can be replaced by any linear transform of the log-spectra. By using MDT-techniques with cepstra, a superior accuracy and robustness relative to their spectral competitors is obtained. The price we pay is that the imputation of the missing data is more complex: the evaluation of a Gaussian now requires the solution of a Non-Negative Least Squares problem. Through the introduction of the ProSpect features [9], the computational load of the cepstral representation is alleviated while their accuracy is maintained. So far, these MDT-techniques could only be used with binary missing data masks. An extension to these MDT-techniques will be presented in this paper such that they can cope with fuzzy masks.

The paper is organized as follows. The missing data techniques with binary masks are first reviewed in section 2 and extended in section 3 for the use with fuzzy masks. In section 4, the performance of these techniques are compared on the Aurora2 connected digit recognition task. Conclusions are given in section 5.

2. MDT FOR BINARY MASKS

The speech recognizer is assumed to have a mainstream HMM-based architecture with Gaussian mixture models (GMM). In the front-end, a low resolution MEL-spectral representation is computed by a filter bank with D channels through windowing, framing, FFT and filter bank integration. Let $\boldsymbol{s}$, $\boldsymbol{n}$ and $\boldsymbol{y}$ denote the vector of D log-MEL spectral features at a certain time frame for the clean speech, the noise and the noisy signal respectively. Ideal binary masks are then obtained by

[*] This work was supported by "Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)".

comparing the log-spectra of clean speech and noise:

$$\boldsymbol{m} = \big(\boldsymbol{s} \geq \boldsymbol{n} - \theta\big)_{0/1} \tag{1}$$

where $(\dots)_{0/1}$ equals 1 (0) when the logical expression inside the brackets holds (does not hold) and θ is a constant threshold. If 1 is assigned to a time-frequency cell, it is dominated by speech, while the mask value 0 indicates that the cell is masked by background noise. The reliable components $\boldsymbol{s}_r$ of the clean speech are approximated by their counterparts in the noisy speech $\boldsymbol{y}_r$, while the unreliable speech components $\boldsymbol{s}_u$ are unknown and have to be estimated.

In the maximum likelihood per Gaussian-based imputation [9], the missing part of $\boldsymbol{s}$ is estimated by minimizing the (negative) log-likelihood Φ for each Gaussian mixture component i over $\boldsymbol{s}$:

$$\begin{aligned} \Phi_i &= \frac{1}{2}(\boldsymbol{s} - \boldsymbol{\mu}_i)' \boldsymbol{P}_i (\boldsymbol{s} - \boldsymbol{\mu}_i) \\ \text{s.t. } & \boldsymbol{s}_r = \boldsymbol{y}_r \text{ and } \boldsymbol{s}_u \leq \boldsymbol{y}_u \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_i$ is the Gaussian mean and $\boldsymbol{P}_i$ is an inverse covariance or precision matrix, both estimated on clean training data. The precision matrix can be expressed in the log-spectral domain or in any other domain that is a linear transformation of log-spectral features. All these variants of MDT have a known symmetric positive semi-definite precision matrix $\boldsymbol{P}_i$. First, we show how to solve (2) in the log-spectral domain, then we discuss the strategy for solving the problem in the ProSpect (or cepstral) domain.

2.1. Log-spectral domain

In most MDT-systems, GMMs with diagonal covariance in the log-spectral domain are used. The optimization problem is then decomposed in D independent problems and each j-th component of the clean speech estimate $\hat{\boldsymbol{s}}_i$ is computed as:

$$\hat{s}_{j,i} = \begin{cases} (1 - m_j)\mu_{j,i} + m_j y_j, & y_j > \mu_{j,i}, \\ y_j, & y_j \leq \mu_{j,i}. \end{cases} \tag{3}$$

2.2. ProSpect domain

Higher accuracies are obtained with GMMs with a diagonal covariance in the cepstral domain, in which case $\boldsymbol{P}_i$ becomes non-diagonal. Imputation then becomes computationally more complex since the estimation of the unreliable part now requires the solution of a Non-Negative Least Square (NNLSQ) problem, e.g. the constrained minimization of a quadratic. In order to reduce the computational load for solving the NNLSQ-problem, the ProSpect features were defined in [9]. Like cepstra, they are computed by a linear transform that has the property of decorrelating the spectral features such that they can be modeled using a GMM with diagonal covariances. While these features can be applied in any speech recognition system, they show a clear benefit in MDT-based recognition in particular since they reduce the computational requirements over cepstral MDT while the accuracy is maintained.

To avoid the costly matrix inversions, the NNLSQ-problem will be solved by the gradient descent method discussed in [9]. Therefore, the search is started from the spectral MDT solution (3). Within each iteration k, these initial values are updated by:

$$\boldsymbol{s}_i^{(k+1)} = \boldsymbol{s}_i^{(k)} - \hat{\alpha} \nabla \boldsymbol{\phi}_i^{(k)} \tag{4}$$

where the step direction $\nabla \boldsymbol{\phi}_i^{(k)}$ is derived from the cost gradient $\nabla \Phi_i^{(k)} = \boldsymbol{P}_i(\boldsymbol{s}_i^{(k)} - \boldsymbol{\mu}_i)$ by zeroing out those components that (i) are labeled as reliable or (ii) where the gradient is negative and the corresponding speech estimate is on the constraint boundary. The optimal step size is given by

$$\alpha = \frac{\nabla \boldsymbol{\phi}_i^{(k)\prime} \nabla \boldsymbol{\phi}_i^{(k)}}{\nabla \boldsymbol{\phi}_i^{(k)\prime} \boldsymbol{P}_i \nabla \boldsymbol{\phi}_i^{(k)}} \tag{5}$$

and is reduced to $\hat{\alpha}$ such that all components of $\boldsymbol{s}_i^{(k+1)}$ satisfy the constraint $\boldsymbol{s}_i^{(k+1)} \leq \boldsymbol{y}$, hence

$$\hat{\alpha} = \min \Big[\min \big(\alpha \nabla \boldsymbol{\phi}_i^{(k)} \, , \, \boldsymbol{y} - \boldsymbol{s}_i^{(k)} \big) ./ \nabla \boldsymbol{\phi}_i^{(k)} \Big] \tag{6}$$

where $./$ denotes the element-wise division. Experiments have shown that the gradient descent method converges in 1 or 2 ($= K$) iterations [9].

Finally, we obtain a clean speech estimate $\hat{\boldsymbol{s}}_i = \boldsymbol{s}_i^{(K)}$ for each Gaussian i, from which we can compute the corresponding likelihood:

$$f(\hat{\boldsymbol{s}}_i | i) = \frac{1}{\sqrt{2\pi}^D \sqrt{|\boldsymbol{P}_i|}} e^{-\frac{1}{2}(\hat{\boldsymbol{s}}_i - \boldsymbol{\mu}_i)' \boldsymbol{P}_i (\hat{\boldsymbol{s}}_i - \boldsymbol{\mu}_i)}. \tag{7}$$

3. MDT FOR FUZZY MASKS

Misclassifications in the binary mask will significantly reduce the recognition rate of the MDT-based recognizer. This can be solved by replacing the hard missing data decision of (1) by a soft-decision strategy. A fuzzy mask vector $\boldsymbol{w}$ can be generated by the approach of [2], e.g. by the substitution of (1) in a sigmoid function:

$$\boldsymbol{w} = \left(\frac{1}{1 + \exp\big(-\rho(\boldsymbol{s} - \boldsymbol{n} - \theta) \big)} \right)_{[0..1]} \tag{8}$$

with slope ρ and where $(\dots)_{[0..1]}$ means that the mask vector now consists of continuous values between 0 and 1. If the value is close to 1, the component has a high probability of being dominated by speech. In order to estimate the clean speech, we need to modify the constrained optimization problem (2) such that it can cope with the probabilistic information provided by the fuzzy masks. Therefore, we need to formulate a new optimization function that should have the ability

that (a) if the mask value is close to 1, the optimal point tends to the observation value, and (b) if the mask value is close to 0, the optimal point tends to a value as close to the Gaussian mean as permitted by the constraint $\boldsymbol{s} \leq \boldsymbol{y}$. Condition (b) is fulfilled if the precision matrix $\boldsymbol{P}_i$ in (2) is replaced by

$$\boldsymbol{Q_i} = \big(\boldsymbol{I} - \boldsymbol{W}\big)^{\frac{1}{2}\prime} \boldsymbol{P}_i \big(\boldsymbol{I} - \boldsymbol{W}\big)^{\frac{1}{2}} \tag{9}$$

where $\boldsymbol{I}$ denotes the $D \times D$ identity matrix and $\boldsymbol{W}$ is the $D \times D$ diagonal matrix with the elements of $\boldsymbol{w}$ on the diagonal. Remark that the matrix multiplications in (9) preserve the symmetric structure of $\boldsymbol{P}_i$. This modification together with the addition of the extra term $\frac{1}{2}(\boldsymbol{s}-\boldsymbol{y})'\boldsymbol{W}(\boldsymbol{s}-\boldsymbol{y})$, will meet condition (a). The proposed optimization problem for fuzzy masks then becomes:

$$\begin{gathered} \Psi_i = \frac{1}{2}(\boldsymbol{s}-\boldsymbol{\mu}_i)'\boldsymbol{Q}_i(\boldsymbol{s}-\boldsymbol{\mu}_i) + \frac{1}{2}(\boldsymbol{s}-\boldsymbol{y})'\boldsymbol{W}(\boldsymbol{s}-\boldsymbol{y}) \\ \text{s.t.} \quad \boldsymbol{s} \leq \boldsymbol{y} \end{gathered} \tag{10}$$

Note that all constraints are now inequality constraints, since there is no evidence for a specific component for being reliable. A description of the solution strategy of (10) in the log-spectral domain and in the ProSpect (or cepstral) domain, can be found in the next subsections.

3.1. Log-spectral domain

In the log-spectral domain, the components of the mean of Ψ_i are given by

$$\bar{\mu}_{j,i} = \frac{(1-w_j)\mu_{j,i}/\sigma^2_{j,i} + w_j y_j}{(1-w_j)/\sigma^2_{j,i} + w_j} \tag{11}$$

with $\sigma^2_{j,i}$ the j-th diagonal element of the log-spectral covariance matrix of Gaussian i. The components of the optimal point $\hat{\boldsymbol{s}}_i$ that minimizes (10) are then found as

$$\hat{s}_{j,i} = \begin{cases} \bar{\mu}_{j,i}, & y_j > \bar{\mu}_{j,i}, \\ y_j, & y_j \leq \bar{\mu}_{j,i}. \end{cases} \tag{12}$$

Remark that for mask values w_j equal to 0 or 1, the solution (12) equals those of (3).

3.2. ProSpect domain

Since (10) is still a constrained minimization problem of a quadratic, a gradient descent method similar to the one explained in section 2.1 is used for solving the problem in the cepstral and ProSpect domain. A good choice to initialize the search is to start from the log-spectral solution (12) or from the point: $\min\big(\boldsymbol{\mu}_i + \boldsymbol{w}(\boldsymbol{y}-\boldsymbol{\mu}_i), \boldsymbol{y}\big)$. This starting point is then iteratively updated by the rule:

$$\boldsymbol{s}_i^{(k+1)} = \boldsymbol{s}_i^{(k)} - \hat{\beta} \nabla \boldsymbol{\psi}_i^{(k)}. \tag{13}$$

The step direction $\nabla \boldsymbol{\psi}_i^{(k)}$ is equal to the cost gradient

$$\nabla \Psi_i^{(k)} = \boldsymbol{Q}_i(\boldsymbol{s}_i^{(k)} - \boldsymbol{\mu}_i) + \boldsymbol{W}(\boldsymbol{s}_i^{(k)} - \boldsymbol{y}) \tag{14}$$

for those components where the corresponding speech estimate lies below the constraint boundary, otherwise the component of $\nabla \boldsymbol{\psi}_i^{(k)}$ will be set to zero. The optimal step size is given by

$$\beta = \frac{\nabla \boldsymbol{\psi}_i^{(k)\prime} \nabla \boldsymbol{\psi}_i^{(k)}}{\nabla \boldsymbol{\psi}_i^{(k)\prime} (\boldsymbol{Q}_i + \boldsymbol{W}) \nabla \boldsymbol{\psi}_i^{(k)}} \tag{15}$$

and is reduced to $\hat{\beta}$ in the same way as was done in section 2.2. Convergence is again reached after 1 or 2 iteration steps. Finally, the cost of each Gaussian i of the acoustic model is obtained by the substitution of $\hat{\boldsymbol{s}}_i$ in (7).

4. EXPERIMENTAL RESULTS

The evaluation of the proposed MDT-based recognizers is done on the Aurora2 TI-Digits speech database, test set A. The acoustic model in the back-end consists of an HMM Gaussian mixture architecture with 16 states per digit and 20 Gaussians per state. The optional inter-word silence is modeled by 1 or 3 states with 36 Gaussians per state, while leading and trailing silence have 3 states. The total number of Gaussians is 3628. The front-end of the MDT-system uses 23-channel MEL filter bank spectra which are transformed to the ProSpect domain for some of the experiments reported below. The static features are compensated with the proposed MDT-methods for binary and fuzzy masks, while dynamic features are left uncompensated since they are more robust to noise.

The accuracy results for the four noise types of the test set are presented in table 1 for ideal (oracle) masks. The binary ideal mask uses (1) as a decision criterion and the fuzzy ideal mask is computed as in (8). The results are compared by solving the optimization problem of (2) and (10) in the log-spectral and ProSpect domain. It can be seen that ProSpect MDT-techniques increase the robustness of the log-spectral MDT-system for both mask versions and that fuzzy masks perform slightly better than binary masks. This can be due to the fact that when speech and noise energy are very close, making a hard decision is not always the best choice.

In a second experiment, we replace the static ideal mask by a real mask computed from the noisy data using harmonicity and SNR information [10]. These masks exploit the strong harmonicity characteristics of voiced speech arising from the presence of the pitch and its harmonics. Hence, the voiced speech can be decomposed into a harmonic signal component which consists of the spectral lines at integer pitch multiples, the remaining spectral lines result in the aperiodic part. The decision criterion of [10] uses the idea that the harmonic part will be dominated by the speech which may lead to poor mask decisions in unvoiced speech segments. As can be seen from table 2, the influence of these masking errors

mask type	SNR (dB)	log-spectral MDT					ProSpect MDT				
		Subw.	Babble	Car	Exhib.	Avg.	Subw.	Babble	Car	Exhib.	Avg.
binary	15	96.41	96.40	97.46	96.20	96.62	98.96	98.73	99.11	98.89	98.92
	10	93.80	94.44	93.92	92.69	93.71	97.70	98.19	97.82	96.73	97.61
	5	86.52	87.55	84.55	82.94	85.39	93.34	95.50	92.07	90.28	92.80
fuzzy	15	96.84	97.01	97.76	96.82	97.11	98.77	98.58	98.78	98.52	98.66
	10	94.50	94.92	95.11	93.77	94.58	97.67	97.94	98.42	97.13	97.79
	5	86.92	88.09	86.10	83.74	86.21	94.66	96.01	94.09	91.92	94.17

Table 1. Recognition accuracy on Aurora2 test set A using ideal masks and log-spectral and ProSpect MDT.

mask type	SNR (dB)	log-spectral MDT					ProSpect MDT				
		Subw.	Babble	Car	Exhib.	Avg.	Subw.	Babble	Car	Exhib.	Avg.
binary	15	95.30	94.01	96.54	95.25	95.28	97.61	97.04	98.12	97.87	97.66
	10	88.39	88.54	91.20	87.84	88.99	94.87	93.29	95.02	93.86	94.26
	5	73.53	69.74	74.47	69.64	71.85	83.14	81.92	81.90	79.51	81.62
fuzzy	15	96.22	95.89	96.78	96.17	96.27	97.61	97.40	98.33	98.09	97.86
	10	91.96	92.50	91.98	90.74	91.80	95.12	94.50	95.47	94.14	94.81
	5	78.11	77.75	75.31	74.08	76.31	85.97	85.37	83.48	81.73	84.14

Table 2. Recognition accuracy on Aurora2 test set A using real masks and log-spectral and ProSpect MDT.

is significantly reduced (especially for 5dB SNR) by casting the decision criterion of [10] in a sigmoid function. These unvoiced regions are now assigned a mask value close to 0.5, e.g. expressing uncertainty, such that more freedom is left to the search algorithm in the recognizer's back-end. Also, the advantage of using MDT-systems with ProSpect features instead of the conventional log-spectral features becomes clear.

5. CONCLUSIONS

We have extended our previous MDT-technique for binary masks such that it can be applied with fuzzy masks. Besides the log-spectral domain, both techniques are applicable for other feature representations, such as cepstra and ProSpects. The experiments showed the advantages of using ProSpect features over the customary log-spectra. We have also demonstrated that the noise robustness of the MDT-based recognizer can further be improved by using fuzzy masks instead of binary masks and this for both ideal (oracle) and real masks. Future work may now exist to exploit probabilistic mask estimation methods in the ProSpect MDT-framework.

6. REFERENCES

[1] P. Renevey and A. Drygajlo, "Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2627–2630.

[2] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, Beijing, China, 2000, pp. 373–376.

[3] M. L. Seltzer, B. Raj, and R.M.. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," in *Speech Comm.*, 2004, vol. 43, no. 4, pp. 379–393.

[4] M. Cooke, Ph. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," in *Speech Comm.*, 2001, vol. 34, pp. 267–285.

[5] J. Barker, M. Cooke, and D.P.W. Ellis, "Decoding speech in the presence of other sources," in *Speech Comm.*, 2005, vol. 45, no. 1, pp. 5–25.

[6] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2837–2840.

[7] B. Raj, M. L. Seltzer, and R.M.. Stern, "Reconstruction of missing features for robust speech recognition," in *Speech Comm.*, 2004, vol. 43, no. 4, pp. 275–296.

[8] H. Van hamme, "Robust Speech Recognition Using Missing Feature Theory in the Cepstral or LDA Domain," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 3089–3092.

[9] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, 2004, pp. 101–104.

[10] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP*, Montreal, Canada, 2004, pp. 213–216.